

21st Nordic Conference of Computational Linguistics (NODALIDA 2017)

NEALT Proceedings Series Volume 29

Gothenburg, Sweden
23 – 24 May 2017

Editor:

Jorg Tiedemann

ISBN: 978-1-5108-4170-3

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2017) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2017)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Regular Papers

<i>Joint UD Parsing of Norwegian Bokmål and Nynorsk</i> Erik Velldal, Lilja Øvrelid and Petter Hohle	1
<i>Replacing OOV Words For Dependency Parsing With Distributional Semantics</i> Prasanth Kolachina, Martin Riedl and Chris Biemann	11
<i>Real-valued Syntactic Word Vectors (RSV) for Greedy Neural Dependency Parsing</i> Ali Basirat and Joakim Nivre	20
<i>Tagging Named Entities in 19th Century and Modern Finnish Newspaper Material with a Finnish Semantic Tagger</i> Kimmo Kettunen and Laura Löfberg	29
<i>Machine Learning for Rhetorical Figure Detection: More Chiasmus with Less Annotation</i> Marie Dubremetz and Joakim Nivre	37
<i>Coreference Resolution for Swedish and German using Distant Supervision</i> Alexander Wallin and Pierre Nugues	46
<i>Aligning phonemes using finite-state methods</i> Kimmo Koskenniemi	56
<i>Twitter Topic Modeling by Tweet Aggregation</i> Asbjørn Steinskog, Jonas Therkelsen and Björn Gambäck	77
<i>A Multilingual Entity Linker Using PageRank and Semantic Graphs</i> Anton Södergren and Pierre Nugues	87
<i>Linear Ensembles of Word Embedding Models</i> Avo Muromägi, Kairit Sirts and Sven Laur	96
<i>Using Pseudowords for Algorithm Comparison: An Evaluation Framework for Graph-based Word Sense Induction</i> Flavio Massimiliano Cecchini, Chris Biemann and Martin Riedl	105
<i>North-Sámi to Finnish rule-based machine translation system</i> Tommi Pirinen, Francis M. Tyers, Trond Trosterud, Ryan Johnson, Kevin Unhammer and Tiina Puolakainen	115
<i>Machine translation with North Saami as a pivot language</i> Lene Antonsen, Ciprian Gerstenberger, Maja Kappfjell, Sandra Nystø Ráhka, Marja-Liisa Olthuis, Trond Trosterud and Francis M. Tyers	123
<i>SWEGRAM –A Web-Based Tool for Automatic Annotation and Analysis of Swedish Texts</i> Jesper Näsman, Beáta Megyesi and Anne Palmér	132
<i>Optimizing a PoS Tagset for Norwegian Dependency Parsing</i> Petter Hohle, Lilja Øvrelid and Erik Velldal	142

<i>Creating register sub-corpora for the Finnish Internet Parsebank</i> Veronika Laippala, Juhani Luotolahti, Aki-Juhani Kyröläinen, Tapio Salakoski and Filip Ginter	152
<i>KILLE: a Framework for Situated Agents for Learning Language Through Interaction</i> Simon Dobnik and Erik de Graaf	162
<i>Data Collection from Persons with Mild Forms of Cognitive Impairment and Healthy Controls - Infrastructure for Classification and Prediction of Dementia</i> Dimitrios Kokkinakis, Kristina Lundholm Fors, Eva Björkner and Arto Nordlund	172
<i>Evaluation of language identification methods using 285 languages</i> Tommi Jauhiainen, Krister Lindén and Heidi Jauhiainen	183
<i>Can We Create a Tool for General Domain Event Analysis?</i> Siim Orasmaa and Heiki-Jaan Kaalep	192
<i>From Treebank to Propbank: A Semantic-Role and VerbNet Corpus for Danish</i> Eckhard Bick	202

Student Papers

<i>Acoustic Model Compression with MAP adaptation</i> Katri Leino and Mikko Kurimo	65
<i>OCR and post-correction of historical Finnish texts</i> Senka Drobac, Pekka Kauppinen and Krister Lindén	70
<i>Mainstreaming August Strindberg with Text Normalization</i> Adam Ek and Sofia Knuutinen	266
<i>Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing</i> Mika Koistinen, Kimmo Kettunen and Tuula Pääkkönen	277
<i>The Effect of Excluding Out of Domain Training Data from Supervised Named-Entity Recognition</i> Adam Persson	289

Short Papers

<i>Cross-lingual Learning of Semantic Textual Similarity with Multilingual Word Representations</i> Johannes Bjerva and Robert Östling	211
<i>Will my auxiliary tagging task help? Estimating Auxiliary Tasks Effectivity in Multi-Task Learning</i> Johannes Bjerva	216
<i>Iconic Locations in Swedish Sign Language: Mapping Form to Meaning with Lexical Databases</i> Carl Börstell and Robert Östling	221
<i>Docforia: A Multilayer Document Model</i> Marcus Klang and Pierre Nugues	226

<i>Finnish resources for evaluating language model semantics</i> Viljami Venekoski and Jouko Vankka	231
<i>Málrómur: A Manually Verified Corpus of Recorded Icelandic Speech</i> Steinþór Steingrímsson, Jón Guðnason, Sigrún Helgadóttir and Eiríkur Rögnvaldsson	237
<i>The Effect of Translationese on Tuning for Statistical Machine Translation</i> Sara Stymne	241
<i>Word vectors, reuse, and replicability: Towards a community repository of large-text resources</i> Murhaf Fares, Andrey Kutuzov, Stephan Oepen and Erik Velldal	271
<i>Redefining Context Windows for Word Embedding Models: An Experimental Study</i> Pierre Lison and Andrey Kutuzov	284
<i>Quote Extraction and Attribution from Norwegian Newspapers</i> Andrew Salway, Paul Meurer, Knut Hofland and Øystein Reigem	293
<i>Wordnet extension via word embeddings: Experiments on the Norwegian Wordnet</i> Heidi Sand, Erik Velldal and Lilja Øvreliid	298
<i>Universal Dependencies for Swedish Sign Language</i> Robert Östling, Carl Börstell, Moa Gärdenfors and Mats Wirén	303

System Demonstration Papers

<i>Multilingwis2 –Explore Your Parallel Corpus</i> Johannes Graën, Dominique Sandoz and Martin Volk	247
<i>A modernised version of the Glossa corpus search system</i> Anders Nøklestad, Kristin Hagen, Janne Bondi Johannessen, Michał Kosek and Joel Priestley .	251
<i>Dep_search: Efficient Search Tool for Large Dependency Parsebanks</i> Juhani Luotolahti, Jenna Kanerva and Filip Ginter	255
<i>Proto-Indo-European Lexicon: The Generative Etymological Dictionary of Indo-European Languages</i> Jouna Pyysalo	259
<i>Tilde MODEL - Multilingual Open Data for EU Languages</i> Roberts Rozis and Raivis Skadiņš	263
<i>Services for text simplification and analysis</i> Johan Falkenjack, Evelina Rennes, Daniel Fahlborg, Vida Johansson and Arne Jönsson	309
<i>Exploring Properties of Intralingual and Interlingual Association Measures Visually</i> Johannes Graën and Christof Bless	314
<i>TALERUM - Learning Danish by Doing Danish</i> Peter Juel Henriksen	318
<i>Cross-Lingual Syntax: Relating Grammatical Framework with Universal Dependencies</i> Aarne Ranta, Prasanth Kolachina and Thomas Hallgren	322

<i>Exploring Treebanks with INESS Search</i>	
Victoria Rosén, Helge Dyvik, Paul Meurer and Koenraad De Smedt	326
<i>A System for Identifying and Exploring Text Repetition in Large Historical Document Corpora</i>	
Aleksi Vesanto, Filip Ginter, Hannu Salmi, Asko Nivala and Tapio Salakoski	330